Classification of dispersive whale calls using a convolutional neural network

Mark Goldwater[1,2], Julien Bonnel[2], Daniel P. Zitterbart[2]

[1]F.W. Olin College of Engineering, [2]Woods Hole Oceanographic Institution

I. Introduction

Background:

- Low-frequency acoustic signals interact with sea surface and seabed
- Can received signal be modeled by a set of dispersive normal modes?
- Measured with a single hydrophone, time-frequency dispersion is used to localize the source and/or estimate the environment

Objective: Architect a convolutional neural network (CNN) to detect multi-modal dispersed gunshots (impulse calls) from Southern right whales

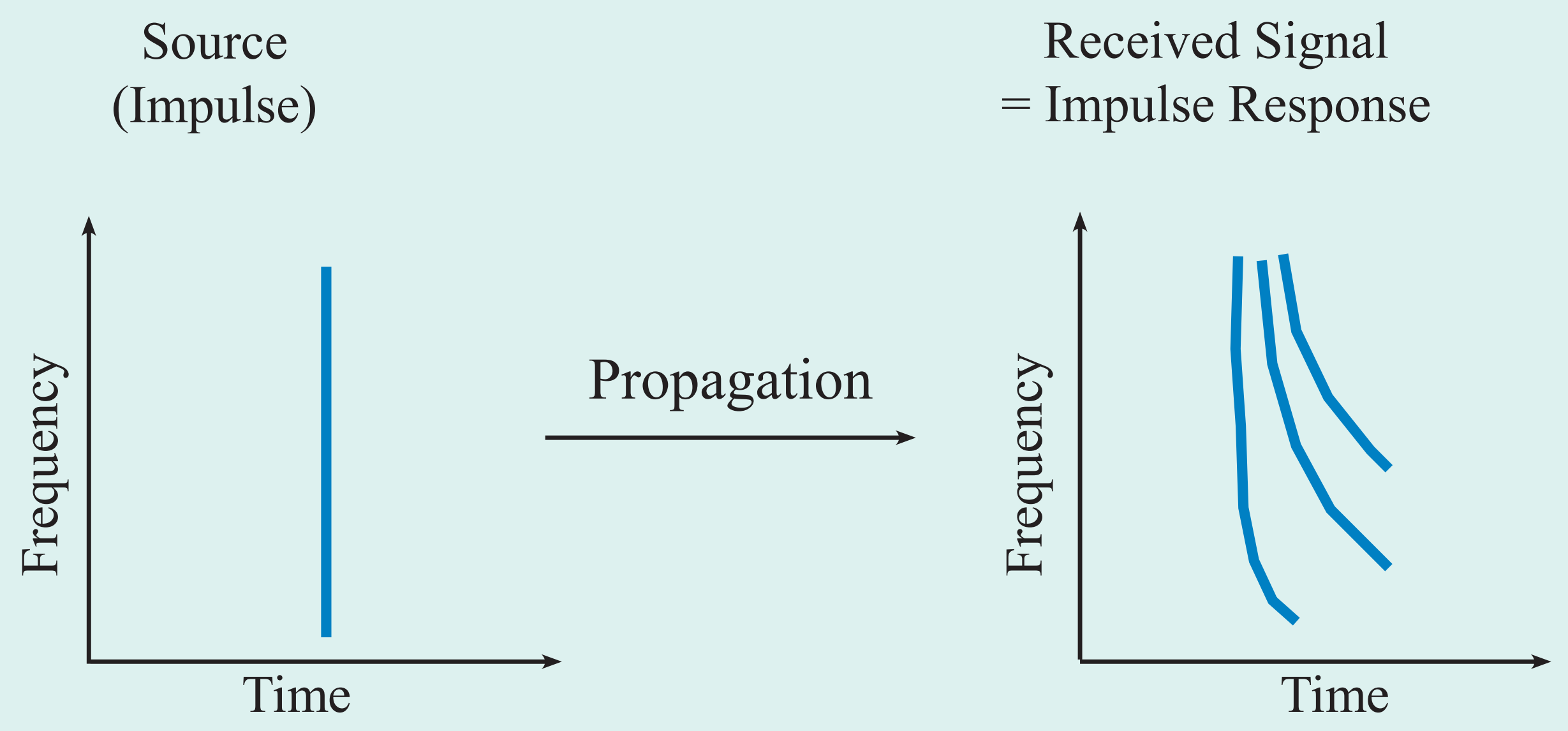

Figure 1: The time-frequency plot on the left shows an impulsive source at t = 0 s. As the source signal propagates, the different modes within the signal begin to disperse because each travels with a different group velocity.

II. Data

- The data was recorded via passive acoustic monitoring over about eight days in late August and early September in Baja de San Antonio, Argentina
- Manually scanned for Southern right whale gunshots and labeled as follows:

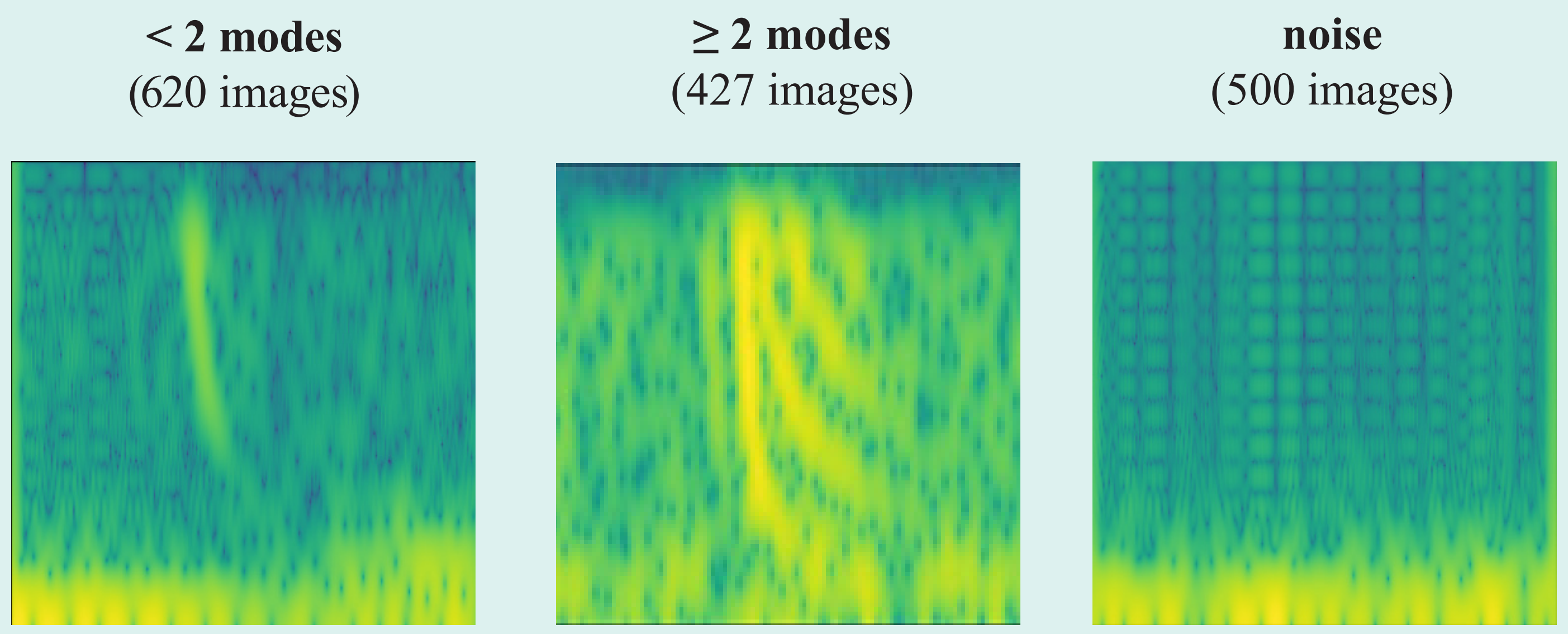

Figure 2: Above are examples from each of the three classes (*less than two modes, at least two modes,* and *noise*) are shown. The CNN was trained to differentiate between these three classes.

- The vast majority of the audio data belongs to the noise class
- Although both classes of recorded calls are useful in ocean acoustics, those which are multi-modal are of particular interest in this research
  - At least two modes required to be present in the recorded call for source ranging and/or environment estimation

III. Acoustic Data Analysis

1. Slide a 0.634-second window forward in the audio file
2. For each window calculate a spectrogram and process with the CNN
3. If identified as multi-modal, save indices of call and CNN confidence for future processing

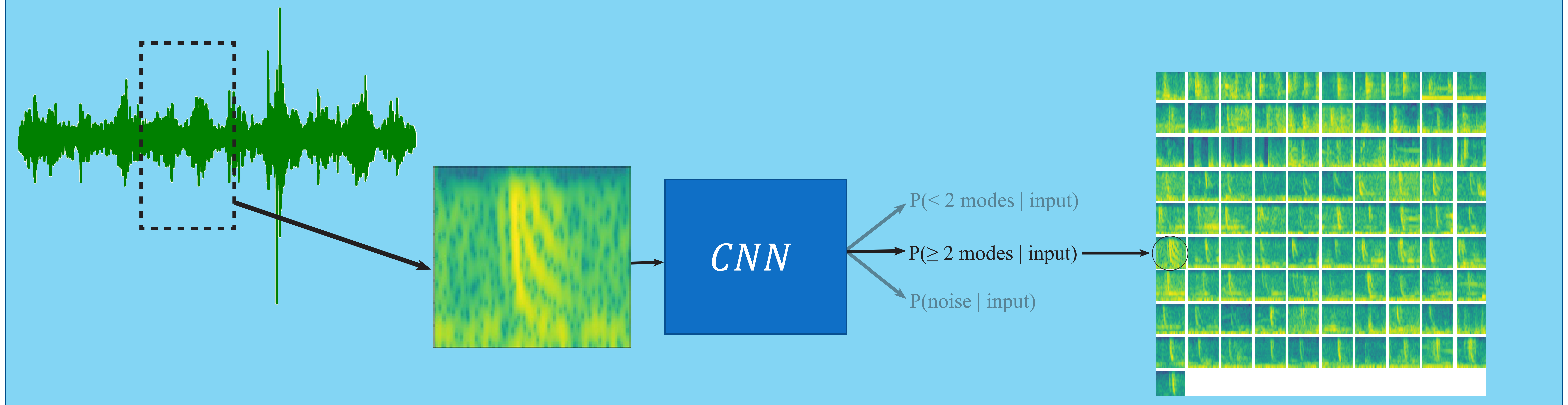

Figure 3: Here, a sample from an audio timeseries is shown undergoing the process used to analyze a large audio file. A 0.634-second clip is extracted and a spectrogram is calculated. This spectrogram is then inputed into the CNN. If the maximum probability in the output layer corresponds to the class of calls which posses at least two modes, the call location and the confidence metrics from the network are saved for future reference and processing. The large timeseries is parsed in 0.634-second adjacent windows such that there is no overlap; however, a small degree of overlap may be useful to detect calls at the boundaries of said partitioning scheme.

IV. Results

- Training performed using 5-fold cross validation
  - Data split into five partitions
  - Five models trained with different validation partition
  - Metrics from each model averaged to evaluate model
- Most mislabeling occurs between the calls with less than two modes and calls with at least two modes
- The intraclass precision for the noise class is almost perfect

Fold	Loss	Accuracy (%)	≥ 2 Modes Precision (%)
1	0.39	89.99	83.15
2	0.32	92.58	82.35
3	0.38	90.29	90.12
4	0.38	89.32	96.82
5	0.36	90.29	87.88
Average	**0.37**	**90.5**	**88.1**

Figure 4: The table here shows the final loss, accuracy, and precision within the *at least two modes* class for each of the five training folds. The last row displays the mean of the metrics for each of the five folds.

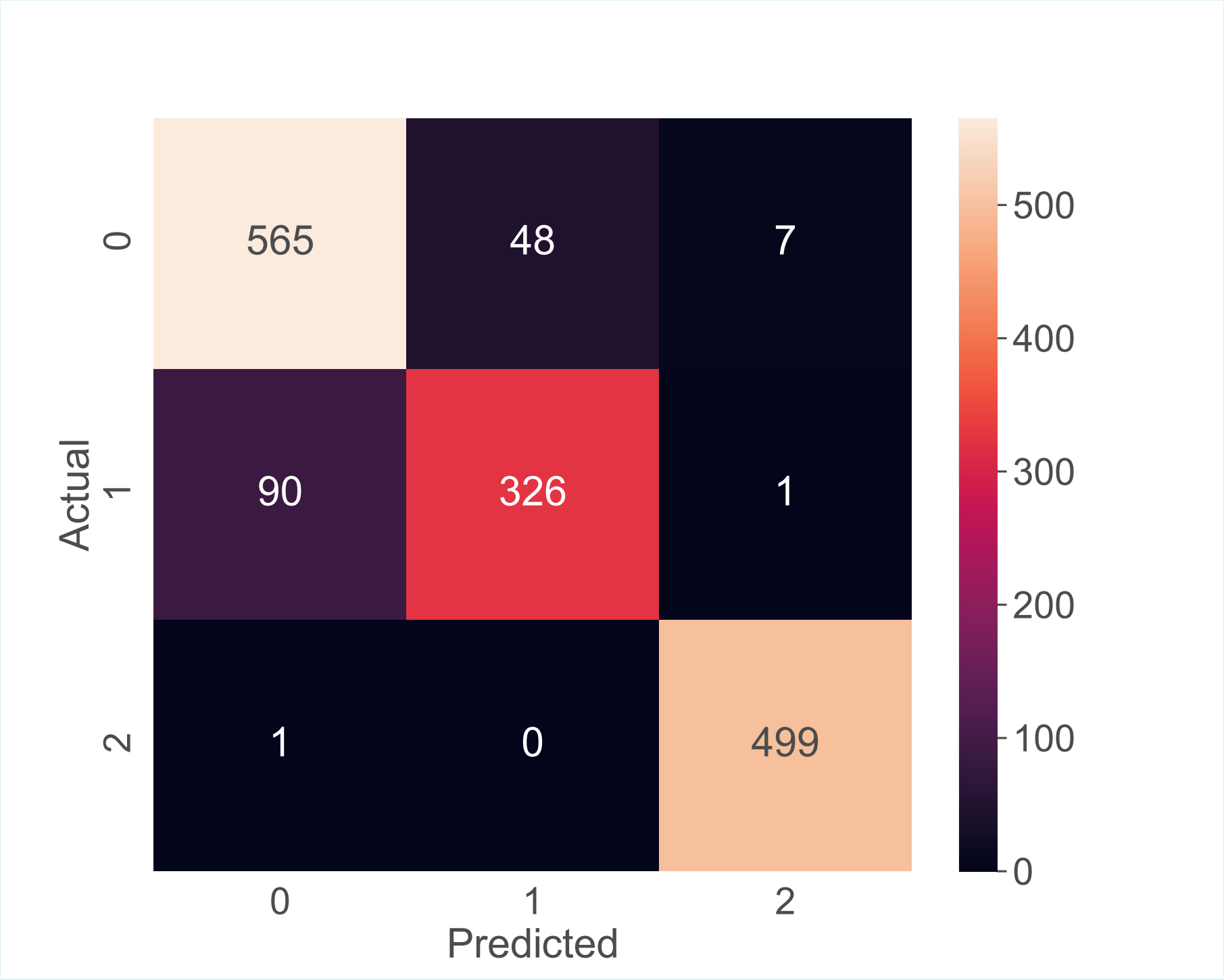

Figure 5: The confusion matrix was calculated by summing the confusion matrices for each of the five folds. Note that most of the mislabeling is among the *less than two modes* and *at least two modes* classes. The *noise* class has almost perfect precision.

V. Conclusion

- CNN is able to quickly identify multi-modal Southern right whale calls with high precision
- Only a few high quality calls are required for source ranging and/or environment estimation
- Modal can significantly reduce the time to process this data using time-frequency dispersion algorithms which require specific input
- In the future, other aspects of the algorithm can benefit from machine learning to reduce manual iteration and enable completely autonomous processing

Acknowledgments

Firstly, I would like to thank my advisors Dr. Julien Bonnel and Dr. Daniel P. Zitterbart for their guidance and support. I would also like to thank the WHOI Summer Student Fellowship who made this research possible.